DOCUMENT RESUME

ED 475 488 EA 032 356

AUTHOR	Greene, Jay P.; Winters, Marcus A.; Forster, Greg
TITLE	Testing High Stakes Tests: Can We Believe the Results of Accountability Tests? Civic Report.
INSTITUTION	Manhattan Inst., New York, NY. Center for Civic Innovation.
REPORT NO	CCI-R-33
PUB DATE	2003-02-00
NOTE	25p.
AVAILABLE FROM	Manhattan Institute for Policy Research, 52 Vanderbilt Avenue, New York, NY 10017. Tel: 212-599-7000; Fax: 212-599-3493; e-mail: holsen@manhattan-institute.org. Web site: http://manhattan-institute.org. For full text: http://www.manhattan-institute.org/cr_33.pdf.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Academic Achievement; *Accountability; Achievement Gains; Educational Assessment; Elementary Secondary Education; Evaluation; *Evaluation Research; *High Stakes Tests; *Mathematics Tests; *Reading Tests; Scores; Standards; *Test Reliability; Test Results
IDENTIFIERS	Colorado; Florida; Illinois; Kansas; Massachusetts; Missouri; No Child Left Behind Act 2001; Ohio; Virginia

ABSTRACT

Many states have implemented high-stakes testing since the enactment of the No Child Left Behind Act of 2001. Yet the question remains whether high-stakes tests effectively measure student proficiency. This report describes a study that compared results on high-stakes tests with results on other standardized tests not used for accountability purposes and thus considered low-stakes tests. Data for the comparisons were gathered from test scores from 5,587 schools in 9 school systems in 8 states. Scores were compared on each test given in the same subject in the same school year. When possible, the results of high-stakes and low-stakes tests given at the same grade levels were also compared. For all the school systems examined in the study, high correlations between score levels on high-stakes and low-stakes tests were found. Also found were some high correlations for year-to-year gains in scores on high-stakes and low-stakes tests. But the correlations of score gains were not as consistently high, and in some places were quite low. The report concludes that stakes of the tests do not distort information about the general level at which students are performing. (Contains 10 tables and 23 references.) (WFA)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Civic Report

No. 33 February 2003

ED 475 488

Testing High Stakes Tests: Can We Believe the Results of Accountability Tests?

Jay P. Greene, Ph.D.
Senior Fellow, Manhattan Institute for Policy Research

Marcus A. Winters
Research Associate, Manhattan Institute for Policy Research

Greg Forster, Ph.D.
Senior Research Associate, Manhattan Institute for Policy Research

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
A Wilson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

CCi
CENTER FOR CIVIC INNOVATION
AT THE MANHATTAN INSTITUTE

EA 032356

BEST COPY AVAILABLE

Executive Summary

Do standardized tests that are used to reward or sanction schools for their academic performance, known as "high stakes" tests, effectively measure student proficiency? Opponents of high stakes testing argue that it encourages schools to "teach to the test," thereby improving results on high stakes tests without improving real learning. Since many states have implemented high stakes testing and it is also central to President Bush's No Child Left Behind Act, this a crucial question to answer.

This report tackles that important policy issue by comparing schools' results on high stakes tests with their results on other standardized tests that are not used for accountability purposes, and thus are "low stakes" tests. Schools have no incentive to manipulate scores on these nationally respected tests, which are administered around the same time as the high stakes tests. If high stakes tests and low stakes tests produce similar results, we can have confidence that the stakes attached to high stakes tests are not distorting test outcomes, and that high stakes test results accurately reflect student achievement.

The report finds that score levels on high stakes tests closely track score levels on other tests, suggesting that high stakes tests provide reliable information on student performance. When a state's high stakes test scores go up, we should have confidence that this represents real improvements in student learning. If schools are "teaching to the test," they are doing so in a way that conveys useful general knowledge as measured by nationally respected low stakes tests. Test score levels are heavily influenced by factors that are outside schools' control, such as student demographics, so some states use year-to-year score gains rather than score levels for accountability purposes. The report's analysis of year-to-year score gains finds that some high stakes tests are less effective than others in measuring schools' effects on student performance.

The report also finds that Florida, which has the nation's most aggressive high stakes testing program, has a very strong correlation between high and low stakes test results on both score levels and year-to-year score gains. This justifies a high level of confidence that Florida's high stakes test is an accurate measure of both student performance and schools' effects on that performance. The case of Florida shows that a properly designed high stakes accountability program can provide schools with an incentive to improve real learning rather than artificially improving test scores.

The report's specific findings are as follows:

- On average in the two states and seven school districts studied, representing 9% of the nation's total public school enrollment, there was a very strong population adjusted average correlation (0.88) between high and low stakes test score levels, and a moderate average correlation (0.45) between the year-to-year score gains on high and low stakes tests. (If the high and low stakes tests produced identical results, the correlation would be 1.00.)
- The state of Florida had by far the strongest correlations, with a 0.96 correlation between high and low stakes test score levels, and a 0.71 correlation between the year-to-year gains on high and low stakes tests.
- The other state studied, Virginia, had a strong 0.77 correlation between test score levels, and a weak correlation of 0.17 between year-to-year score gains.
- The Chicago school district had a strong correlation of 0.88 between test score levels, and no correlation (-0.02) between year-to-year score gains.
- The Boston school district had a strong correlation of 0.75 between test score levels, and a moderate correlation of 0.27 between year-to-year score gains.

- The Toledo school district had a strong correlation of 0.79 between test score levels, and a weak correlation of 0.14 between year-to-year score gains.
- The Fairfield, Ohio, school district had a moderate correlation of 0.49 between test score levels, and a moderate negative correlation of -0.56 between year-to-year score gains.
- The Blue Valley, Kansas, school district had a moderate correlation of 0.53 between test score levels, and a weak correlation of 0.12 between year-to-year score gains.
- The Columbia, Missouri, school district had a strong correlation of 0.82 between test score levels, and a weak negative correlation of -0.14 between year-to-year score gains.
- The Fountain Fort Carson, Colorado, school district had a moderate correlation of 0.35 between test score levels, and a weak correlation of 0.15 between year-to-year score gains.

About the Authors

Jay P. Greene is a Senior Fellow at the Manhattan Institute for Policy Research where he conducts research and writes about education policy. He has conducted evaluations of school choice and accountability programs in Florida, Charlotte, Milwaukee, Cleveland, and San Antonio. He also recently published a report and a number of articles on the role of funding incentives in special education enrollment increases.

His research was cited four times in the Supreme Court's opinions in the landmark *Zelman* v. *Simmons-Harris* case on school vouchers. His articles have appeared in policy journals, such as *The Public Interest*, *City Journal*, and *Education Next*, in academic journals, such as *The Georgetown Public Policy Review*, *Education and Urban Society*, and *The British Journal of Political Science*, as well as in major newspapers, such as the *Wall Street Journal* and *Christian Science Monitor*.

Greene has been a professor of government at the University of Texas at Austin and the University of Houston. He received his B.A. in history from Tufts University in 1988 and his Ph.D. from the Government Department at Harvard University in 1995. He lives with his wife and three children in Weston, Florida.

Marcus Winters is a Research Associate at the Manhattan Institute's Education Research Office where he studies and writes on education policy. He recently graduated from Ohio University with a B.A. in political science, for which he received departmental honors, and a minor in economics.

Greg Forster is a Senior Research Associate at the Manhattan Institute's Education Research Office. He is the co-author of several education studies and op-ed articles. He received a Ph.D. with distinction in Political Science from Yale University in May 2002, and his B.A. from the University of Virginia, where he double-majored in Political and Social Thought and Rhetoric and Communications Studies, in 1995.

Acknowledgements

The authors would like to thank Chester E. Finn, Jr. and Rick Hess, who reviewed this manuscript and provided valuable comments and suggestions. The authors would also like to thank the state and local education officials who helped make test score data available for the study.

Table of Contents

Introduction 1
A Variety of Testing Policies 1
Previous Research 2
Method 4
Results 7
Conclusion 8
Appendix 9
Table 1: Average Correlations 9
Table 2: Florida 10
Table 3: Virginia 11
Table 4: Chicago, IL 11
Table 5: Boston, MA 12
Table 6: Toledo, OH 12
Table 7: Blue Valley, KS 13
Table 8: Columbia, MO 13
Table 9: Fairfield, OH 14
Table 10: Fountain Fort Carson, CO 14
References 15
Endnotes 17

Testing High Stakes Tests: Can We Believe the Results of Accountability Tests?

Introduction

"High stakes" testing, the use of standardized tests to reward or sanction schools for their academic performance, is among the most contentious issues in education policy. As the centerpiece of President Bush's No Child Left Behind Act, it is also among the most prominent education reform strategies. The idea behind it is that rewarding or sanctioning schools for their performance provides schools with incentives necessary to improve academic achievement.

But what if schools respond to the incentives of high stakes testing by developing ways to improve results on the high stakes tests without actually improving real learning? This is the principal objection raised by opponents of high stakes testing. Opponents contend that schools will "teach to the test," or cheat even more directly by manipulating the test answer sheets. The concern raised by opponents is that high stakes testing causes schools to teach skills or adopt policies that are only useful for passing the high stakes tests and are not more generally useful in helping prepare students for later life.

Whether the high stakes of high stakes testing are in fact motivating schools to manipulate results without actually improving real student achievement is a question that can be investigated empirically. By comparing results from high stakes tests with results from other standardized tests administered around the same time, we can determine whether the high stakes associated with high stakes tests are distorting test results. If high stakes tests produce results that are similar to the results of other tests where there are no incentives to manipulate scores, which we might call "low stakes" tests, then we can have confidence that the high stakes do not themselves distort the outcomes. If, on the other hand, high stakes tests produce results that are *not* similar to the results of low stakes tests, then we should be concerned that schools have managed to produce results on high stakes tests that are inaccurate reflections of actual student achievement.

This report investigates the validity of high stakes testing by comparing the results of high and low stakes tests administered to students around the same time in two states and in seven school districts nationwide. The states and districts examined contain 9% of all public school students in the country. We find that scores on high and low stakes tests generally produce results that are similar to each other. The population adjusted average correlation between high and low stakes test results in all the school systems we examined was 0.88, which is a very strong correlation (if the high and low stakes tests produced identical results, the correlation would be 1.00).

We also find that year-to-year improvement on high stakes testing is strongly correlated with year-to-year improvement on low stakes standardized tests in some places, but weakly correlated in others. The population adjusted average correlation between year-to-year gain on high stakes tests and year-to-year gain on low stakes tests in all the school systems we examined was 0.45, which is a moderately strong correlation. But the correlation between year-to-year gains on Florida's high and low stakes tests was extremely high, 0.71, while the correlation in other locations was considerably lower. These analyses lead us to conclude that well-designed high stakes accountability systems can and do produce reliable measures of student progress, as they appear to have done in Florida, but we can have less confidence that other states' high stakes tests are as well designed and administered as Florida's.

A Variety of Testing Policies

There is considerable diversity in testing policies nationwide. States and school districts around the country vary in the types of tests they use, the number of subjects they test, the grades in which they

administer the tests, and the seriousness of the sanctions or rewards they attach to test results. Some states, such as Minnesota, report scores on state-mandated tests to the public in order to shame school districts into performing better; other states, such as Ohio and Massachusetts, require students to pass the state exam before receiving a high school diploma. Chicago public school students must perform well on the Iowa Test of Basic Skills in specified grades in order to be promoted to the next grade, even though neither the test nor the sanction is required by the state of Illinois.

Perhaps the nation's most aggressive test-based accountability measure is Florida's A+ program. Florida uses results on the Florida Comprehensive Assessment Test (FCAT) to hold students accountable by requiring all students to pass the 3rd grade administration of the exam before moving to the 4th grade, and by withholding diplomas from students who have not passed all sections of the 10th grade administration of the exam. It also holds schools and districts accountable by using FCAT results to grade schools from A to F on school report cards that are very widely publicized and scrutinized. However, what really makes Florida's program stand out is that the state holds schools and districts accountable for their students' performance on FCAT by offering vouchers to all students in schools that have earned an F on their report cards in any two of the previous four years. These chronically failing schools face the possibility of the ultimate consequence -- they could lose their students and the state funding that accompanies them.

Two states, Florida and Virginia, and several school districts gave their students both a high stakes test and a commercially-designed low stakes test during the school year. The low stakes tests are used to assess how well students are doing compared to national norms and to decide what curriculum changes should be implemented to better serve students. Since parents and school officials see the results of the tests and use them for their own purposes, it would be incorrect to say that there are no stakes attached to them at all. However, the stakes attached to these tests are small enough that schools have little or no incentive to manipulate the results in the way that some fear high stakes tests may be manipulated. Thus a student's performance on a low stakes test is most likely free from potential distortion.

Previous Research

Several objections have been raised against using standardized testing for accountability purposes. Most concerns about high stakes testing revolve around the adverse incentives created by the tests. Some have worried that pressures to produce gains in test scores have led to poor test designs or questionable revisions in test designs that exaggerate student achievement (for example, see Koretz and Barron 1998 on Kentucky's test; Haney 2000 on Texas' test; and Haney et al 1999 on Massachusetts' test). Others have written that instead of teaching generally useful skills, teachers are teaching skills that are unique only to a particular test (for example, see Amrein and Berliner 2002; Klein et al 2000; McNeil and Valenzuela 2000; Haney 2000; and Koretz and Barron 1998). Still others have directly questioned the integrity of those administering and scoring the high stakes tests, suggesting that cheating has produced much of the claimed rise in student achievement on such exams (for example, see Cizek 2001; Dewan 1999; Hoff 1999; and Lawton 1996).

Most of these criticisms fail to withstand scrutiny. Much of the research done in this area has been largely theoretical, anecdotal, or limited to one or another particular state test. For example, Linda McNeil and Angela Valenzuela's critique of the validity of high stakes testing lacks an analysis of data (see McNeil and Valenzuela 2000). Instead, their arguments are based largely on theoretical expectations and anecdotal reports from teachers, whose resentment of high stakes testing for depriving them of autonomy may cloud their assessments of the effectiveness of testing policies. Their reports of cases in which high stakes tests were manipulated are intriguing, but they do not present evidence on whether these practices are sufficiently widespread to fundamentally distort testing results.

Other researchers have compared high stakes test results to results on other tests, as we do in this study. Prior research in this area, however, has failed to use tests that accurately mirror the population of students taking the high stakes test or the level of knowledge needed to pass the state mandated exam.

Amrein and Berliner find a weak relationship between the adoption of high stakes tests and improvement in other test indicators, such as NAEP,

SAT, ACT, and AP results (see Amrein and Berliner 2002).[1] Koretz and Barron find that Kentucky's high stakes test results show increases that are not similarly found in the state's NAEP results (see Koretz and Barron 1998). Klein et al similarly claim that gains on the Texas high stakes test appear to be larger than are shown by NAEP (see Klein, Hamilton, McCaffrey, and Stecher 2000).

Comparing state-mandated high stakes tests with college entrance and AP exams is misleading because the college-oriented exams are primarily taken by the best high school students, who represent a minority of all students. Though the percentage of students taking these exams has increased to the point that test-takers now include more than the most elite students, they still are not taken by all students, and this hinders their usefulness for assessing the validity of near-universally administered high stakes tests. Only a third of all high school students take the SAT, and even fewer take the ACT or AP. Furthermore, college-oriented tests tell us nothing about the academic progress of the student population that high stakes testing is most intended to benefit: low performing students in underserved communities. In addition, because these tests are intended only for college bound students they test a higher level of knowledge than most high stakes tests, which are used to make sure students have the most basic knowledge necessary to earn a diploma. Any discrepancy between the results of college-oriented tests and high stakes tests could be attributable to the difference in the populations taking these tests and the different sets of skills they demand.

Comparisons between high stakes tests and NAEP are more meaningful than comparisons to college-oriented tests, though NAEP-based analyses also fall short of the mark. NAEP is administered infrequently and only to certain grades. Any weak correlation between NAEP and high stakes tests could be attributable to such factors. When tests are not administered around the same time, or are not administered to the same students, their results are less likely to track each other. This will soon change with the new, more frequent NAEP testing schedule required under the No Child Left Behind Act – although NAEP will also become a high stakes test under No Child Left Behind, so its usefulness for evaluating other tests may not be improved.

Rather than focusing on statewide outcomes, like NAEP or college-oriented exam results, Haney uses classroom grades to assess the validity of Texas' high stakes test. He finds a weak correlation between Texas high stakes results and classroom grades, from which he concludes that the Texas high stakes test results lack credibility (see Haney 2000). However, it is far more likely that classroom grades lack credibility. Classroom grades are highly subjective and inconsistently assigned, and are thus likely to be misleading indicators of student progress (see Barnes and Finn 2002). To support this suspicion of classroom grades, Figlio and Lucas, 2001 correlated school grades with scores in Florida on the state's high stakes test and found that teacher given grades were inflated (see Figlio and Lucas 2001).

There have also been a number of responses to these critiques of state testing validity. For example, Hanushek and Phelps have written a series of methodological critiques of the work by Haney and Klein (see Hanushek 2001 and Phelps 2001). Hanushek points out that Klein's finding of stronger gains on the Texas state test than on NAEP should come as no surprise given that Texas school curricula are more closely aligned with the Texas test than with NAEP (see Hanushek 2001). Phelps takes Haney and Klein to task for a variety of errors, alleging (for example) that Haney used incorrect NAEP figures on exemption rates in Texas and that Klein failed to note more significant progress on NAEP by Texas students because of excessive disaggregation of scores (see Phelps 2000).

Other analyses, such as those by Grissmer, et al, and Greene, also contradict Haney and Klein's results. Contrary to Haney and Klein, Grissmer and Greene find that Texas made exceptional gains on the NAEP as state-level test results were increasing dramatically (see Grissmer, Flanagan, Kawata, and Williamson 2000; and Greene 2000). Unfortunately, our inability to correlate individual-level or school-level performance on the NAEP and the Texas test, as well as the infrequent administration of NAEP, prevent any clear resolution of this dispute.

This report differs from other analyses in that it focuses on the comparison of school-level results on high stakes tests and commercially-designed low stakes tests. By focusing on school-level results we are comparing test results from the same or similar

students, reducing the danger that population differences may hinder the comparison. Examining school-level results also allows for a more precise correlation of the different kinds of test results than is possible by looking only at state-level results, which provide fewer observations for analysis. In addition, school-level analyses are especially appropriate because in most cases the accountability consequences of high stakes test results are applied at the school level. By comparing school-level scores on high stakes and low stakes tests, this study attempts to find where, if anywhere, we can believe high stakes test results. If we see that high stakes and low stakes tests produce similar results, we have reason to believe that results on the high stakes test were not affected by any of the adverse incentives tied to the test.

Method

The first step in conducting this study was to locate states and school districts that administer both high stakes and low stakes tests. We examined information available on each state's Department of Education website about their testing programs, and contacted by phone states whose information was unclear. A test was considered high stakes if any of the following depended upon it: student promotion or graduation, accreditation, funding cuts, teacher bonuses, a widely publicized school grading or ranking system, or state assumption of at least some school responsibilities. We found two states, Florida and Virginia, that administered both a high stakes test and a low stakes test.[2] Test scores in Florida were available on the Florida Department of Education's website, and we were able to obtain scores from Virginia through a data request.

We next attempted to find individual school districts that also administered both high stakes and low stakes tests. We first investigated the 58 member districts of the Council for Great City Schools, which includes many of the largest school districts in the nation. Next, through internet searches, we looked for other school districts that administer multiple tests. After locating several of these districts, we contacted them by phone and interviewed education staffers about the different types of tests the districts administered.

Because we were forced to rely on internet searches and non-systematic phone interviews to find school districts that gave both high and low-stakes tests, our search was certainly not exhaustive.[3] But the two states and seven school districts included in this study, which did administer both high and low stakes tests, contain approximately 9% of all public school students in the United States and a significantly higher percentage of all students who take a high stakes test. We therefore have reason to believe that our results provide evidence on the general validity of high stakes testing nationwide.

In each of the states and school districts we studied, we compared scores on each test given in the same subject and in the same school year. In total, we examined test scores from 5,587 schools in nine school systems. When possible, we also compared the results of high and low stakes tests given at the same grade levels. We were not able to do this for all the school systems we studied, however, because several districts give their low stakes tests at different grade levels from those that take their high stakes tests. When a high or low stakes test was administered in multiple grade levels of the same school level (elementary, middle, or high school), we took an average of the tests for that school level. Though this method does not directly compare test scores for the same students on both tests, the use of school-level scores does reflect the same method used in most accountability programs.

Because we sometimes had to compute an average test score for a school, and because scores were reported in different ways (percentiles, scale scores, percent passing, etc.), we standardized scores from each separate test administration by converting them into what are technically known as "z-score" results. To standardize the test scores into z-scores, we subtracted the score a school received on the test administration by the average score on that administration throughout the district/state. We then divided that number by the standard deviation of the test administration. The standardized test score is therefore equal to the number of standard deviations each school's result is from the sample average.

In school systems with accountability programs, there is debate over how to evaluate test results. School systems evaluate test results in one of two ways: either they look at the actual average test score in each school or they look at how much each school

improved its test scores from one year to another. Each method has its advantages and disadvantages. Looking at score levels tells us whether or not students are performing academically at an acceptable level, but it does not isolate the influence of schools from other factors that contribute to student performance, such as family and community factors. Looking at year-to-year score gains is a "value added" approach, telling us how much educational value each school added to its students in each year.

For the school systems we studied, we computed the correlation between high and low stakes test results for both the score level and the year-to-year gain in scores. We found the year-to-year gain scores for each test by subtracting the standardized score on the test administration in one year from the standardized score on the test administration in the next year. For example, in Florida we subtracted each school's standardized score on the 4th grade reading FCAT test in 2000 from the same school's standardized score on the 4th grade reading FCAT in 2001. This showed us whether a school was either gaining or losing ground on the test.

We used a Pearson's correlation to measure how similar the results from the high and low stakes tests were, both in terms of score levels and in terms of the year-to-year gain in scores. For example, for score levels we measured the correlation between the high stakes FCAT 3rd grade reading test in 2001 and the low stakes Standford-9 3rd grade reading test in 2001. Similarly, for year-to-year score gains we measured the correlation between the 2000–2001 score gain on the FCAT and the 2000–2001 score gain on the Standford-9.[4]

Where there is a high correlation between high and low stakes test results, we conclude that the high stakes of the high stakes test do not distort test results, and where there is a low correlation we have significantly less confidence in the validity of the high stakes test results.[5]

There are many factors that could explain a low correlation between high and low stakes test results. One possibility would be that the high stakes test is poorly designed, such that schools can successfully target their teaching on the skills required for the high stakes test without also conveying a more comprehensive set of skills that would be measured by other standardized tests. It is also possible that the implementation of high stakes tests in some school systems could be poorly executed. Administering high stakes tests in only a few grades may allow schools to reallocate their best teachers to those grades, creating false improvements that are not reflected in the low stakes test results from other grades. The security of high stakes tests could also be compromised, such that teachers and administrators could teach the specific items needed to answer the questions on the high stakes test without at the same time teaching a broader set of skills covered by the low stakes standardized test. It is even possible that in some places teachers and administrators have been able to manipulate the high stakes test answers to inflate the apparent performance of students on the high stakes test.

More benign explanations for weak correlations between high and low stakes test results are also available. When we analyze year-to-year gains in the test scores, we face the problem of having to measure student performance twice, thus introducing more measurement error. Weak correlations could also partially be explained by the fact that the score gains we examine do not track a cohort of the same students over time. Such data are not available, forcing us to compute the difference in scores between one year's students against the previous year's students in the same grade. While this could suppress the correlation of gain scores, it is important to note that our method is comparable to the method of evaluation used in virtually all state high stakes accountability systems that have any kind of value-added measurement. In addition, if a school as a whole is in fact improving, we would expect to observe similar improvement on high and low stakes tests when comparing the same grades over time.

Correlations between results on high and low stakes tests could also be reduced to some extent by differences in the material covered by different tests. High stakes tests are generally geared to a particular state or local curriculum, while low stakes tests are generally national. But this can be no more than a partial explanation of differences in test results. There is no reason to believe that the set of skills students should be expected to acquire in a particular school system would differ dramatically from the skills covered by nationally-respected standardized

tests. Students in Virginia need to be able to perform arithmetic and understand what they read just like students in other places, especially if students in Virginia hope to attend colleges or find employment in other places.

If low correlations between results on high and low stakes tests are attributable to differences between the skills required for the two tests, we might reasonably worry that the high stakes test is not guiding educators to cover the appropriate academic material. It might be the case that the high stakes test is too narrowly drawn, such that it does not effectively require teachers to convey to their students a broad set of generally useful skills. The low stakes tests used in the school systems we studied are all nationally respected tests that are generally acknowledged to measure whether or not students have successfully achieved just this kind of broad skill learning, so if the high stakes test results in these systems do not correlate with their low stakes test results, this may be an indication that poorly-designed high stakes tests are failing to cover a broad set of skills. On the other hand, if their high stakes test results are strongly correlated with their results on low stakes tests that are nationally respected as measurements of broad skill learning, this would give us a high degree of confidence that the high stakes tests are indeed testing a broad set of generally useful skills and not just a narrow set of skills needed only to pass the test itself.

Interpretation of our results is made somewhat problematic because we cannot know with absolute certainty the extent to which factors other than school quality influence test score levels. Family background, population demographics, and other factors are known to have a significant effect on students' level of achievement on tests, but we have no way of knowing how large this effect is. To an unknown extent, score level correlations reflect other factors in addition to the reliability of the high stakes test. However, the higher the correlation between score levels on high and low stakes tests, the less we have reason to believe that poor test design or implementation undermines the reliability of high stakes test results. Furthermore, where a high correlation between year-to-year score gains accompanies a high correlation between score levels, we can be very confident that the high stakes test is reliably measuring school quality because family and demographic factors have no significant effect on score gains. On the other hand, even where score level correlations are high, score gain correlations could be low if student background factors are causing the high score level correlations.

No doubt some will object that a high correlation between high and low stakes test scores does not support the credibility of high stakes tests because they do not believe that low stakes standardized tests are any better than high stakes standardized tests as a measure of student achievement. Some may question whether students put forth the necessary effort on a test with no real consequences tied to their scores. This argument would prove true if we find low correlations between the tests on the score levels. If a large number of students randomly fill in answers on the low stakes test, then that randomness will produce low correlations with the high stakes tests, on which the students surely gave their best effort. But where we find high correlations on the score levels we have confidence that students gave comparable effort on the two tests.

Others may object entirely to the use of standardized testing to assess student performance. To those readers, no evidence would be sufficient to support the credibility of high stakes testing, because they are fundamentally opposed to the notion that academic achievement can be systematically measured and analyzed by standardized tests. The difficulty with this premise is that it leads to educational nihilism. If academic achievement cannot be systematically measured, we cannot ever know whether or not students in general are making progress, nor can we ever know in general whether schools are helping, hurting, or having no effect on student progress. If we cannot know these things, then we cannot identify which educational techniques are likely to be effective or which policy interventions are likely to be desirable.

This study begins with a different premise: that achievement is measurable. Its purpose is to address the reasonable concern that measurements of achievement are distorted by the accountability incentives that are designed to spur improvement in achievement. By comparing scores on tests where there may be incentives to distort the results with scores on tests where there are almost no incentives to distort the results, we are able to isolate the extent

to which the incentives of high stakes testing are in fact distorting information on student achievement.

Results

For all the school systems examined in our study, we generally found high correlations between score levels on high and low stakes tests.[6] We also found some high correlations for year-to-year gains in scores on high and low stakes tests, but the correlations of score gains were not as consistently high, and in some places were quite low.

This greater variation on score gain correlations might be partially explained by the increased measurement error involved in calculating score gains as opposed to score levels. It is also possible that high stakes tests provide less credible measures of student progress in some school systems than in others. In places where high stakes tests are poorly designed (such that teaching to the test is an effective strategy for boosting performance on the high stakes test without also conveying useful skills that are captured by the low stakes test) or where the security of tests has been compromised (such that teachers can teach the exact items to be included in the high stakes test, or help students cheat during the test administration), the correlations between score gains on high and low stakes tests may be quite low. The high correlations between score level results on high and low stakes tests do not rule out these possibilities because, to an unknown extent, score levels reflect family and demographic factors in addition to school quality. However, despite this, the high correlations between score level results do justify a moderate level of confidence in the reliability of these systems' high stakes tests.

Perhaps the most intriguing results we found came from the state of Florida. We might expect the especially large magnitude of the stakes associated with Florida's high stakes test to make it highly vulnerable to adverse responses because of the incentives created by high stakes testing. It was in Florida, however, that we found the highest correlations between high and low stakes test results, for both score levels in each given year and for the year-to-year score gains.

Florida's high stakes test, the FCAT, produced score levels that correlated with the score levels of the low stakes Stanford-9 standardized test across all grade levels and subjects at 0.96. If the two tests had produced identical results, the correlation would have been 1.00. The year-to-year score gains on the FCAT correlated with the year-to-year score gains on the Stanford-9 at 0.71. (See the Appendix Tables for the average correlations as well as the separate correlations between each test, in each subject, and for each test administration.) Both of these correlations are very strong, suggesting that the high and low stakes tests produced very similar information about student achievement and progress. Because the high stakes FCAT produces results very similar to those from the low stakes Stanford-9, we can be confident that the high stakes associated with the FCAT did not distort its results. If teachers were "teaching to the test" on the FCAT, they were teaching generally useful skills that were also reflected in the results of the Stanford-9, a nationally respected standardized test.

In other school systems we found very strong correlations between score levels for high and low stakes test results in each given year, but relatively weak or even negative correlations between the year-to-year score gains on the two types of tests. For example, in Virginia the correlation between score levels on the state's high stakes Standards of Learning test (SOL) and the low stakes Stanford-9 was 0.77, but the correlation between the year-to-year score gains on these two tests was only 0.17. Similarly, in Boston the correlation between the level of the high stakes Massachusetts Comprehensive Assessment System (MCAS) and the low stakes Stanford-9 was 0.75, but the correlation on the gain in scores between these two tests was a moderate 0.27. In Toledo, Ohio, the correlation between the level of the high and low stakes tests was 0.79, while the correlation between the score gains on the same tests was only 0.14.

In Chicago, the Iowa Test of Basic Skills (ITBS) is administered as a high stakes test in some grades and a low stakes test in other grades. The correlation between score levels on the high and low stakes administrations of this test is a very strong 0.88. But the year-to-year score gain in the results of the ITBS in high stakes grades is totally uncorrelated (-0.02) with the year-to-year score gain from the same test given in grades where the stakes are low. Similarly, in Columbia, Missouri, the high correlation (0.82) of

score levels on the high and low stakes tests is accompanied by a weak negative correlation (-0.14) between the year-to-year score gain on the two types of tests.

In some school systems even the level of results on high and low stakes tests correlate only moderately well. In Blue Valley, Kansas, the high and low stakes tests produce score levels that correlate at 0.53 and score gains that correlate at only 0.12. In Fairfield, Ohio, the score levels on the high and low stakes tests correlate at 0.49, while, oddly, the year-to-year score gains have a moderate negative correlation of -0.56. In Fountain Fort Carson, Colorado, the score level correlation is only 0.35, while the score gain correlation is an even weaker 0.15.

Conclusion

The finding that high and low stakes tests produce very similar score level results tells us that the stakes of the tests do not distort information about the general level at which students are performing. If high stakes testing is only being used to assure that students can perform at certain academic levels, then the results of those high stakes tests appear to be reliable policy tools. The generally strong correlations between score levels on high and low stakes tests in all the school systems we examined suggest that teaching to the test, cheating, or other manipulations are not causing high stakes tests to produce results that look very different from tests where there are no incentives for distortion.

But policymakers have increasingly recognized that score level test results are strongly influenced by a variety of factors outside of a school system's control. These include student family background, family income, and community factors. If policymakers want to isolate the difference that schools and educators make in student progress, they need to look at year-to-year score gains, or "value-added" measures, as part of a high stakes accountability system.

Florida has incorporated value-added measures into its high stakes testing and accountability system, and the evidence shows that Florida has designed and implemented a high stakes testing system where the year-to-year score gains on the high stakes test correspond very closely with year-to-year score gains on standardized tests where there are no incentives to manipulate the results. This strong correlation suggests that the value-added results produced by Florida's high stakes testing system provide credible information about the influence schools have on student progress.

In all of the other school systems we examined, however, the correlations between score gains on high and low stakes tests are much weaker. We cannot be completely confident that those high stakes tests provide accurate information about school influence over student progress. However, the consistently high correlations we found between score levels on high and low stakes tests does justify a moderate level of confidence in the reliability of those high stakes tests.

Our examination of school systems containing 9% of all public school students shows that accountability systems that use high stakes tests can, in fact, be designed to produce credible results that are not distorted by teaching to the test, cheating, or other manipulations of the testing system. We know this because we have observed at least one statewide system, Florida's, where high stakes have not distorted information either about the level of student performance or the value that schools add to their year-to-year progress. In other school systems we have found that high stakes tests produce very credible information on the level of student performance and somewhat credible information on the academic progress of students over time. Further research is needed to identify ways in which other school systems might modify their practices to produce results more like those in Florida.

Appendix

Table 1
Average Correlations

	Level Scores	Gain Scores
Florida	0.96	0.71
Virginia	0.77	0.17
Chicago, IL	0.88	-0.02
Boston, MA	0.75	0.27
Toledo, OH	0.79	0.14
Blue Valley, KS	0.53	0.12
Columbia, MO	0.82	-0.14
Fairfield, OH	0.49	-0.56
Fountain Fort Carson, CO	0.35	0.15
Total Average (Weighted by Population)	**0.88**	**0.45**

Table 2
Florida

Level Score Correlations		Gain Score Correlations	
Grade 3 Math, 2001	0.97	Grade 3 Math, 2002–2001	0.81
Grade 3 Math, 2002	0.96	Grade 3 Reading, 2002–2001	0.88
Grade 3 Reading, 2001	0.98	Grade 4 Math, 2002–2001	0.76
Grade 3 Reading, 2002	0.97	Grade 4 Reading, 2001–2000	0.73
Grade 4 Math, 2001	0.97	Grade 4 Reading, 2002–2001	0.77
Grade 4 Math, 2002	0.96	Grade 5 Math, 2001–2000	0.76
Grade 4 Reading, 2000	0.96	Grade 5 Math, 2002–2001	0.79
Grade 4 Reading, 2001	0.96	Grade 5 Reading, 2002–2001	0.80
Grade 4 Reading, 2002	0.95	Grade 6 Math, 2002–2001	0.76
Grade 5 Math, 2000	0.95	Grade 6 Reading, 2002–2001	0.78
Grade 5 Math, 2001	0.96	Grade 7 Math, 2002–2001	0.74
Grade 5 Math, 2002	0.95	Grade 7 Reading, 2002–2001	0.66
Grade 5 Reading, 2001	0.98	Grade 8 Math, 2001–2000	0.60
Grade 5 Reading, 2002	0.98	Grade 8 Math, 2002–2001	0.75
Grade 6 Math, 2001	0.97	Grade 8 Reading, 2001–2000	0.65
Grade 6 Math, 2002	0.97	Grade 8 Reading, 2002–2001	0.73
Grade 6 Reading, 2001	0.97	Grade 9 Math, 2002–2001	0.66
Grade 6 Reading, 2002	0.98	Grade 9 reading, 2002–2001	0.67
Grade 7 Math, 2002	0.97	Grade 10 Math, 2001–2000	0.48
Grade 7 Reading, 2001	0.97	Grade 10 Math, 2002–2001	0.68
Grade 7 Reading, 2001	0.97	Grade 10 Reading, 2001–2000	0.43
Grade 7 Reading, 2002	0.97	Grade 10 Reading, 2002–2001	0.70
Grade 8 Math, 2000	0.95		
Grade 8 Math, 2001	0.95	**Average**	**0.71**
Grade 8 Math, 2002	0.97		
Grade 8 Reading, 2000	0.96		
Grade 8 Reading, 2001	0.97		
Grade 8 Reading, 2002	0.97		
Grade 9 Math, 2001	0.97		
Grade 9 Math, 2002	0.96		
Grade 9 Reading, 2001	0.97		
Grade 9 Reading, 2002	0.97		
Grade 10 Math, 2000	0.90		
Grade 10 Math, 2001	0.95		
Grade 10 Math, 2002	0.94		
Grade 10 Reading, 2000	0.89		
Grade 10 Reading, 2001	0.97		
Grade 10 Reading, 2002	0.96		
Average	**0.96**		

High Stakes: Florida Comprensive Assessment Test
Low Stakes: Stanford-9

Table 3
Virginia

Level Score Correlations		Gain Score Correlations	
Elementary Math, 1998	0.84	Elementary Math, 1999–1998	0.40
Elementary Math, 1999	0.85	Elementary Math, 2000–1999	0.34
Elementary Math, 2000	0.80	Elementary Math, 2001–2000	0.21
Elementary Math, 2001	0.75	Middle School Math, 1999–1998	-0.08
Middle School Math, 1998	0.83	Middle School Math, 2000–1999	0.12
Middle School Math, 1999	0.70	Middle School Math, 2001–2000	0.04
Middle School Math, 2000	0.75		
Middle School Math, 2001	0.68	**Average**	**0.17**
Average	**0.77**		

High Stakes: Standards of Learning
Low Stakes: Stanford-9

Table 4
Chicago, IL

Level Score Correlations		Gain Score Correlations	
Elementary Math, 1997	0.85	Elementary Math, 1998–1997	0.03
Elementary Math, 1998	0.85	Elementary Math, 1999–1998	0.04
Elementary Math, 1999	0.82	Elementary Math, 2000–1999	0.00
Elementary Math, 2000	0.81	Elementary Math, 2001–2000	0.06
Elementary Math, 2001	0.86	Elementary Reading, 1998–1997	0.03
Elementary Reading, 1997	0.88	Elemantary Reading, 1999–1998	0.02
Elementary Reading, 1998	0.89	Elementary Reading, 2000–1999	0.08
Elementary Reading, 1999	0.85	Elementary Reading, 2001–2000	0.04
Elementary Reading, 2000	0.86	Middle School Math, 1998–1997	-0.13
Elementary Reading, 2001	0.89	Middle School Math, 1999–1998	-0.12
Middle School Math, 1997	0.92	Middle School Math, 2000–1999	-0.02
Middle School Math, 1998	0.90	Middle School Math, 2001-2000	-0.08
Middle School Math, 1999	0.90	Middle School Reading, 1998–1997	-0.10
Middle School Math, 2000	0.88	Middle School Reading, 1999–1998	-0.05
Middle School Math, 2001	0.89	Middle School Reading, 2000–1999	-0.09
Middle School Reading, 1997	0.93	Middle School Reading, 2001–2000	-0.09
Middle School Reading, 1998	0.91		
Middle School Reading, 1999	0.91	**Average**	**-0.02**
Middle School Reading, 2000	0.88		
Middle School Reading, 2001	0.88		
Average	**0.88**		

High Stakes: Iowa Test of Basic Skills, grades 3,6,8
Low Stakes: Iowa Test of Basic Skills, grades 4,5,7

Table 5
Boston, MA

Level Score Correlations		Gain Score Correlations	
Elementary Math, 1998	0.50	Elementary Math 1999–1998	0.15
Elementary Math, 1999	0.53	Elementary Math, 2000–1999	-0.11
Elementary Math, 2000	0.62	Middle School Math, 1999–1998	0.28
Middle School Math, 1998	0.88	Middle School Math, 2000–1999	0.52
Middle School Math, 1999	0.88	High School Math, 1999–1998	0.36
Middle School Math, 2000	0.89	High School Math, 2000–1999	0.39
High School Math, 1998	0.96		
High School Math, 1999	0.95	**Average**	**0.27**
High School Math, 2000	0.57		
Average	**0.75**		

High Stakes: Massachusetts Comprensive Assessment System
Low Stakes: Stanford-9

Table 6
Toledo, OH

Level Score Correlations		Gain Score Correlations	
Elementary Math, 1999	0.87	Elementary Math, 2000–1999	0.16
Elementary Math, 2000	0.75	Elementary Math, 2001–2000	0.15
Elementary Math, 2001	0.85	Elementary Math, 2002–2001	0.26
Elementary Math, 2002	0.81	Elementary Reading, 1999–1998	-0.05
Elementary Reading, 1998	0.78	Elementary Reading, 2000–1999	0.12
Elementary Reading, 1999	0.88	Elementary Reading, 2001–2000	0.08
Elementary Reading, 2000	0.87	Elementary Reading, 2002–2001	0.31
Elementary Reading, 2001	0.84	Elementary Science, 2002–2001	-0.07
Elementary Reading, 2002	0.87	Middle School Reading, 1999–1998	-0.38
Elementary Science, 2001	0.80	Middle School Reading, 2000–1999	0.43
Elementary Science, 2002	0.41	Middle School Math, 2000–1999	0.51
Middle School Reading, 1998	0.99	Middle School Science, 2000–1999	0.20
Middle School Reading, 1999	0.68	Middle School Social Studies, 2000–1999	0.07
Middle School Reading, 2000	0.82		
Middle School Reading, 2001	0.55	**Average**	**0.14**
Middle School Math, 1999	0.89		
Middle School Math, 2000	0.80		
Middle School Math, 2002	0.84		
Middle School Science, 1999	0.82		
Middle School Science, 2000	0.89		
Middle School Science, 2002	0.90		
Middle School Social Studies, 1999	0.85		
Middle School Social Studies, 2000	0.58		
Middle School Social Studies, 2002	0.64		
Average	**0.79**		

High Stakes: Ohio Proficiency
Low Stakes: Stanford-9

Table 7
Blue Valley, KS

Level Score Correlations		Gain Score Correlations	
Elementary Math, 2000	0.27	Elementary Math, 2001–2000	0.19
Elementary Math, 2001	0.52	Elementary Reading, 2001–2000	0.26
Elementary Reading, 2000	0.44	Middle School Math, 2001–2000	0.11
Elementary Reading, 2001	0.24	Middle School Reading, 2001–2000	0.26
Elementary Science, 2001	0.44	High School Reading, 2001–2000	-0.21
Middle School Math, 2000	-0.03		
Middle School Math, 2001	0.21	**Average**	**0.12**
Middle School Reading, 2000	0.62		
Middle School Reading, 2001	0.14		
Middle School Science, 2001	0.62		
Middle School Social Studies, 2001	0.72		
High School Reading, 2000	0.88		
High School Reading, 2001	0.99		
High School Science, 2001	0.96		
High School Social Studies, 2001	0.95		
Average	**0.53**		

High Stakes: Kansas Assessment
Low Stakes: Iowa Test of Basic Skills

Table 8
Columbia, MO

Level Score Correlations		Gain Score Correlations	
Elementary Math, 1999	0.90	Elementary Math, 2000–1999	0.15
Elementary Math, 2000	0.71	Elementary Math, 2001–2000	-0.43
Elementary Math, 2001	0.86		
		Average	**-0.14**
Average	**0.82**		

High Stakes: Missouri Assessment Program

Low Stakes: Iowa Test of Basic Skills, 1998–99, 1999–2000
Low Stakes: Stanford-9, 2000–2001

Table 9
Fairfield, OH

Level Score Correlations		Gain Score Correlations	
Elementary Reading, 2001	0.01	Elementary Average Reading, 2002–001	-0.56
Elementary Math, 2001	0.90		
Elementary Science, 2001	0.82		
Elementary Social Studies, 2001	0.39		
Elementary Reading, 2002	0.33		
Average	**0.49**		

High Stakes: Ohio Proficiency
Low Stakes: Terra Nova

Table 10
Fountain Fort Carson, CO

Level Score Correlations		Gain Score Correlations	
Elementary Reading, 1999	-0.12	Elementary Average Reading, 2000–1999	0.15
Elementary Math, 1999	0.25		
Elementary Reading, 2000	0.35		
Average	**0.35**		

High Stakes: Colorado Student Assessment Program
Low Stakes: Iowa Test of Basic Skills

References

Audrey L. Amrein and David C. Berliner, "High-Stakes Testing, Uncertainty, and Student Learning," *Education Policy Analysis Archives*, Volume 10 Number 18, March 28, 2002. http://epaa.asu.edu/epaa/v10n18/

Christopher Barnes and Chester E. Finn, "What Do Teachers Teach? A Survey of America's Forth and Eighth Grade Teachers" September, 2002.

Gregory J. Cizek, "Cheating to the Test," *Education Matters*, vol. 1, no. 1, Spring 2001. http://educationnext.org/2001sp/40.html

Shaila Dewan, "The Fix Is In. Are educators cheating on TAAS? Is anyone going to stop them?" *The Houston Press*, February 25, 1999. http://www.houstonpress.com/issues/1999-02-25/feature.html/1/index.html

Education Week's Quality Counts 2002. http://www.edweek.org/sreports/qc02/

David N. Figlio and Maurice E. Lucas, "Do High Grading Standards Effect Student Performance?" December, 2001

Jay P. Greene, "Technicalities: Value added analysis is a crucial tool in the accountability toolbox – despite its flaws," *Education Next*, Forthcoming Summer 2002.

Jay P. Greene, "The Looming Shadow: Can the Threat of Vouchers Persuade a Public School to Turn Itself Around? The Case of Florida Suggests Yes," *Education Next*, Winter 2001. http://educationnext.org/20014/76.html

Jay P. Greene, "An Evaluation of the Florida A-Plus Accountability and School Choice Program" Florida State University, The Manhattan Institute, and The Harvard Program on Education Policy and Governance, February 2001. http://www.manhattan-institute.org/html/cr_aplus.htm

Jay P. Greene, "The Texas School Miracle is for Real," *City Journal*, Summer 2000. http://www.city-journal.org/html/10_3_the_texas_school.html

Jay P. Greene and Greg Forster, "Burning High Stakes Tests at the Stake," *The Education Gadfly*, volume 3, number 1, January 8, 2003. http://www.edexcellence.net/gadfly/v03/gadfly01.html#jaygreg1

Jay P. Greene and Greg Forster, "Effects of Funding Incentives on Special Education Enrollment," Civic Report 32, The Manhattan Institute, December 2002. http://www.manhattan-institute.org/html/cr_32.htm

David W. Grissmer, Ann Flanagan, Jennifer Kawata, and Stephanie Williamson, "Improving Student Achievement: What State NAEP Test Scores Tell Us," Rand Report, July 25, 2000. http://www.rand.org/publications/MR/MR924/

Walt Haney, "The Myth of the Texas Miracle in Education," *Education Policy Analysis Archives*, Volume 8 Number 41, August 19, 2000. http://epaa.asu.edu/epaa/v8n41/index.html

Walt Haney, Clarke Fowler, Anne Wheelock, Damian Bebell and Nicole Malec, "Less Truth Than Error? An independent study of the Massachusetts Teacher Tests," *Education Policy Analysis Archives*, Volume 7 Number 4, February 11, 1999. http://epaa.asu.edu/epaa/v7n4/

Eric A. Hanushek, "Deconstructing RAND," *Education Matters*, vol. 1, no. 1, Spring 2001. http://educationnext.org/2001sp/65.html

David J. Hoff, "N.Y.C. Probe Levels Test-Cheating Charges," *Education Week*, December 15, 1999. http://www.edweek.org/ew/ewstory.cfm?slug=16cheat.h19

Stephen P. Klein, Laura S. Hamilton, Daniel F. McCaffrey, and Brian M. Stecher, "What Do Test Scores in Texas Tell Us?," Rand Report, October 24, 2000. http://www.rand.org/publications/IP/IP202/

Daniel M. Koretz and Sheila I. Barron, "The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS)," Rand Report, 1998. http://www.rand.org/publications/MR/MR1014/#contents

Millicent Lawton, "Alleged Tampering Underscores Pitfalls of Testing," *Education Week*, November 13, 1996. http://www.edweek.org/ew/vol-16/11cheat.h16

Robert D. Mason, et al, *Statistical Techniques in Business and Economics, 10th Edition*, McGraw-Hill, 1999.
Linda McNeil and Angela Valenzuela, "The Harmful Impact Of The TAAS System Of Testing In Texas: Beneath The Accountability Rhetoric," The Harvard Civil Rights Project, January 6, 2000. http://www.law.harvard.edu/civilrights/conferences/testing98/drafts/mcneil_valenzuela.html
Richard P. Phelps, "Test Bashing Series," *EducationNews.org*, 2000.http://www.educationnews.org/test_bashing_series_by_richard_p.htm

ENDNOTES

1. For a criticue specifically of the Amrein and Berliner study see Greene and Forster 2003.

2. A number of states and school districts administer a standardized test in addition to the state criterion reference test, but many of those standardized tests had high stakes attached to the results. For example, Houston and Dallas, Texas, Arizona, and California all administered multiple tests to their students but all tests had high stakes. We could not include those states or school districts in our sample.

3. Because school level test scores are public information and usually covered under state freedom of information laws we might have expected obtaining the scores to have been relatively easy. Unfortunately, we encountered numerous delays and refusals from school officials. Some school districts were very helpful with their test score information and provided us with the necessary data. Other school districts, however, were less helpful and in some cases were downright hostile. The Maynard, Massachusetts school district, for instance, refused to give us the data. We spoke directly to the Assistant Superintendent of the district, who said she was in charge of testing. She informed us that she would not release the test score information because she was "philosophically opposed" to our study. We are unaware how her philosophical opposition trumps public information laws, but since we had neither the time nor the resources to pursue the matter in the courts she was successful in denying us her test score information. The Maynard, Massachusetts case was by far the most blatant obstruction we faced while attempting to obtain the necessary test scores, but some other districts were reluctant to provide the information until we informed them that they were legally required to do so. We found this rather disturbing considering that public schools claim their transparency as one of their greatest virtues. In performing this study, at least, we certainly did not find public schools to be transparent.

4. Our method can be illustrated by using Virginia's administration of the high stakes SOL and the low stakes Standford-9 elementary math tests in 2000 as an example. In this year, Virginia gave the SOL to students in the 3rd and 5th grade, and gave the Stanford-9 to 4th graders. We averaged the 3rd and 5th grade scores on the SOL test to get a single school score on that test.

We next standardized the scores on each of the tests. The SOL was reported as mean scaled scores and the Stanford-9 scores were reported as mean percentiles. We calculated both the average school score on each test and the standard deviation on each test administration. On the SOL the average school mean scaled score was 431.93 and the standard deviation was 39.31. On the Stanford-9 the average school percentile was 57.93 and the standard deviation was 15.24. For each school we subtracted the average school score on the test from that individual school's score on the test and divided the resulting number by the standard deviation. So for Chincoteague Elementary School, which posted a 60 percentile score on the Stanford-9 the calculation was thus:

$$\frac{60 - 57.93}{15.24} = .14$$

After standardizing scores for every school in the state on each of the two test administrations in question, Stanford-9 4th grade math, 2000, and SOL elementary average math, 2000, we then correlated the standard scores on the two tests. In this instance we find a correlation of .80. This high correlation leads us to conclude that in this case the stakes of the tests had no effect on the results of the tests.

We then found and correlated the gain scores for each test. Building off our example, we subtracted the standardized scores on the 1999 administration of the tests from the standardized scores in the 2000 administration of the tests to find the gain or loss the school made on the test in the year. In our example school, this meant a .01 standard score gain on the Stanford-9 and a .10 standard score gain on the SOL. We calculated the gain scores for each school in the state and correlated the results. In this example we found a correlation of .34, a moderate correlation between the two tests.

Next we combined the standardized scores of the test by grade, while keeping them separated by year and subject and correlated the results. In our example this meant combining all 2000 administrations of the Stanford-9 math test (elementary, middle and high school scores) and doing the same for the SOL math 2000 test and correlating the results. In this example we found a high correlation of .77. We then repeated this combining and correlating for the difference scores. In our example we found that the difference between the 2000 and 1999 standardized scores on the SOL in all grades correlated with the difference between the 2000 and 1999 standardized scores on the Stanford-9 in all grades at a level of .29, a moderate correlation.

5. There is one distortion that might be caused by the incentives created by the high stakes of high stakes tests that this method cannot detect: if school systems are excluding low-performing students from the testing pool altogether, such as by labeling them as disabled or non-English speaking, a high correlation between scores on high and low stakes tests would not reveal it. However, the research that has been done so far on exclusion from high stakes testing gives us no good reason to believe that this is occurring to a significant extent. Most studies of this phenomenon are methodologically suspect, and those that are not have found no significant relationship between high stakes testing and testing exclusion (for a full discussion, see Greene and Forster 2002).

6. It is conventional to describe correlations between .75 and 1 as strong correlations, correlations between .25 and .75 as moderate correlations, and correlations between 0 and .25 as weak correlations (Mason, et al., 1999).

EXECUTIVE DIRECTOR
Henry Olsen

ADVISORY BOARD
Stephen Goldsmith, Chairman
Mayor Jerry Brown
Mayor John O. Norquist
Mayor Martin O'Malley
Mayor Rick Baker

FELLOWS
William D. Eggers
Jay P. Greene
Byron R. Johnson
George L. Kelling
Edmund J. McMahon
Peter D. Salins

The Center for Civic Innovation's (CCI) purpose is to improve the quality of life in cities by shaping public policy and enriching public discourse on urban issues.

CCI sponsors the publication of books like The Entrepreneurial City: A How-To Handbook for Urban Innovators, which contains brief essays from America's leading mayors explaining how they improved their cities' quality of life; Stephen Goldsmith's The Twenty-First Century City, which provides a blueprint for getting America's cities back in shape; and George Kelling's and Catherine Coles' Fixing Broken Windows, which explores the theory widely credited with reducing the rate of crime in New York and other cities. CCI also hosts conferences, publishes studies, and holds luncheon forums where prominent local and national leaders are given opportunities to present their views on critical urban issues. *Cities on a Hill*, CCI's newsletter, highlights the ongoing work of innovative mayors across the country.

The Manhattan Institute is a 501(C)(3) nonprofit organization. Contributions are tax-deductible to the fullest extent of the law. EIN #13-2912529

MANHATTAN INSTITUTE FOR POLICY RESEARCH
52 Vanderbilt Avenue • New York, NY 10017
www.manhattan-institute.org

Non-Profit
Organization
US Postage
PAID
Permit 04001
New York, NY

BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: "Testing High Stakes Tests: Can We Believe the Results of Accountability Tests?	
Author(s): Jay P. Greene, Marcus A. Winters, Greg Forster	
Corporate Source: Manhattan Institute	Publication Date: February 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: [signature]	Printed Name/Position/Title: Alexander Wilson
Organization/Address: The Manhattan Institute	Telephone: 212-599-7000 FAX: 212-599-3494
	E-Mail Address: awilson@manhattan-institute.org Date: 4/10/03

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Urban Education
Box 40, Teachers College, Columbia University
New York, NY 10027

Telephone: 212-678-3433
Toll Free: 800-601-4868
Fax: 212-678-4012

WWW: http://eric-web.tc.columbia.edu

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)